

Capturing Common and Individual Components in fMRI Data by Discriminative Dictionary Learning

Krishna Dontaraju, Seung-Jun Kim, Mohammad Akhonda, Tülay Adalı
Dept. of Computer Science & Electrical Engineering
University of Maryland, Baltimore County
Baltimore, MD 21250
{krishna5, sjkim, mo32, adali}@umbc.edu

Abstract—Data-driven analysis for functional magnetic resonance imaging (fMRI) data has played an important role for uncovering salient brain functional networks that are shared across multiple subjects. On the other hand, recent fMRI studies indicate that there is significant and consistent heterogeneity present across different subject groups and individuals. While independent component analysis (ICA) has been a major tool to perform data-driven analysis of fMRI data, dictionary learning (DL) approaches are increasingly receiving attention due to their modeling capability and flexibility. In this work, a supervised DL framework is employed to capitalize on the available class labels and capture not only the commonly shared components across the population, but also the unique components that contribute to discrimination. A systematic comparison with conventional ICA is performed based on real fMRI data consisting of healthy controls and patients with schizophrenia.

I. INTRODUCTION

Analysis of functional magnetic resonance imaging (fMRI) data has played a major role in understanding the brain function. Various data-driven methods have been developed based on latent variable models and matrix and tensor factorization approaches, with the large-scale analysis receiving much attention recently [1]. However, most of the approaches have focused on characterizing common activities shared across the population. On the other hand, it has been recognized that significant and consistent heterogeneity exists across different subgroups and individuals [2], [3].

Blind source separation (BSS) approaches such as independent component analysis (ICA) have been the major data-driven fMRI data analysis methods. Maximizing the statistical independence between latent factors, ICA proved effective for uncovering non-overlapping brain regions without prior knowledge on the temporal structures in the fMRI data [4], [5], and has been recently extended to multiple datasets using independent vector analysis [6].

Dictionary learning (DL) aims to extract a set of representative bases from data by capitalizing on the notion of sparsity [7]. The approach can not only learn an overcomplete basis that entails a powerful union-of-subspaces model, but is also quite flexible in that various prior information can be accommodated into the cost function. The DL approach was applied to fMRI data analysis recently with promising results [8]–[11]. In these works, the fMRI data, arranged in matrices, were factorized to yield sparse spatial activations in the voxel domain. This was done in a purely unsupervised learning framework, however, where the reconstruction fidelity of the factorization was the main objective of training.

DL can be used for supervised learning as well [12], [13]. The idea is to train a dictionary that is not only effective for reconstruction but also for predicting the available labels. In the context of fMRI data

This work was supported in part by NSF grant 1631838 and a UMBC seed grant.

analysis, we are not necessarily interested in the prediction performance itself, but rather in revealing both the common components that are shared across the subjects, and the individual components that are specific to different subjects or subject groups.

The shared and the subject-specific dictionaries were jointly learned for fMRI data with an incoherence term for the subject-specific dictionaries in [14]. The problem of learning a shared dictionary as well as the class-specific dictionaries was formulated by incorporating the low-rank constraints on the shared dictionary in addition to the incoherence terms for class-specific dictionaries [15]. However, having incoherent dictionaries for individual classes may not be desirable in such scenarios where the underlying discriminative sources have overlapping subspaces. For example, the fMRI response for patients with schizophrenia and healthy controls have significantly overlapping responses [16]. Here, having class-specific incoherent dictionaries can split the spatial components, rendering neurological interpretation difficult.

Aiming at learning both the common and the distinctive dictionaries that are also readily interpretable, we formulate a supervised DL problem with a common dictionary and a single discriminative dictionary that is used for all classes. The discriminability is encouraged by Fisher’s discriminant criterion as in [13], [15]. A systematic comparison is done between the conventional ICA method and the proposed DL methods in terms of capturing the commonly shared components across the subjects as well as the distinctive components characteristics of the subgroups. A real dataset comprising 121 schizophrenic subjects and 150 healthy controls is analyzed.

The remainder of the paper is organized as follows. Sec. II provides the DL problem formulations, and Sec. III derives the algorithms for solving the proposed formulations. Sec. IV presents the setup for analyzing the fMRI data. Sec. V provides the results from the analysis. Finally, Sec. VI offers conclusions.

II. PROBLEM FORMULATION

A. Discriminative Dictionary Learning

Given the data matrix $\mathbf{X} := [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N] \in \mathbb{R}^{V \times N}$, DL is an unsupervised learning technique that obtains a dictionary $\mathbf{D} \in \mathbb{R}^{V \times K}$ such that each datum (each column of \mathbf{X}) can be well represented by a linear combination of a small number of columns in \mathbf{D} . This can be achieved by solving e.g. [17]

$$\min_{\mathbf{D}, \mathbf{Z}} \frac{1}{2} \|\mathbf{X} - \mathbf{D}\mathbf{Z}\|_F^2 + \lambda \|\mathbf{Z}\|_1 \quad (1)$$

where $\|\cdot\|_F$ is the Frobenius norm, $\|\mathbf{Z}\|_1 := \sum_{n=1}^N \|\mathbf{z}_n\|_1$, with \mathbf{z}_n being the n -th column of \mathbf{Z} , promotes sparsity in \mathbf{Z} , and $\lambda > 0$ adjusts the sparsity level. It is also understood that \mathbf{D} satisfies some normalization constraints, such as constraining the norm of

each column to be no more than unity, in order to mitigate the scaling ambiguity of the bifactor decomposition.

When the data belong to C different classes, one can develop a supervised learning method that computes a discriminative dictionary for the classification task [12]. One way to accomplish this is to incorporate a cost function related to Fisher's discriminant criterion; see also [13]. Suppose that the n -th datum \mathbf{x}_n belongs to class $c(n)$ and let \mathcal{N}_c represent the set of indices of the class c samples, i.e., $\mathcal{N}_c := \{n : c(n) = c\}$. Also let N_c be the cardinality of \mathcal{N}_c with $\sum_{c=1}^C N_c = N$. Then, upon defining the class mean and the total mean vectors as

$$\mathbf{m}_c := \frac{1}{N_c} \sum_{n \in \mathcal{N}_c} \mathbf{z}_n \text{ and } \mathbf{m} := \frac{1}{N} \sum_{n=1}^N \mathbf{z}_n \quad (2)$$

respectively, the so-called within-class scatter and the between-class scatter matrices are defined as

$$\mathbf{S}_W(\mathbf{Z}) := \sum_{c=1}^C \sum_{n \in \mathcal{N}_c} (\mathbf{z}_n - \mathbf{m}_c)(\mathbf{z}_n - \mathbf{m}_c)^T \quad (3)$$

$$\mathbf{S}_B(\mathbf{Z}) := \sum_{c=1}^C N_c (\mathbf{m}_c - \mathbf{m})(\mathbf{m}_c - \mathbf{m})^T \quad (4)$$

respectively, where T denotes transposition. Thus, a discriminative DL formulation is

$$\min_{\mathbf{D}, \mathbf{Z}} \frac{1}{2} \|\mathbf{X} - \mathbf{D}\mathbf{Z}\|_F^2 + \lambda \|\mathbf{Z}\|_1 + \frac{\mu}{2} f(\mathbf{Z}) \quad (5)$$

where Fisher's criterion is captured by

$$f(\mathbf{Z}) = \text{tr}(\mathbf{S}_W(\mathbf{Z})) - \text{tr}(\mathbf{S}_B(\mathbf{Z})) + \|\mathbf{Z}\|_F^2. \quad (6)$$

The presence of the last term $\|\mathbf{Z}\|_F^2$ renders f convex with respect to \mathbf{Z} [18]. To see this, first define an N -by- N matrix \mathbf{H}_1 whose (i, j) -entry is equal to $1/N_c$ if both \mathbf{x}_i and \mathbf{x}_j belong to the same class, where c is the class to which they belong, and equal to 0 otherwise. Define also an N -by- N matrix \mathbf{H}_2 , whose entries are all equal to $1/N$. Then, upon further defining

$$\begin{aligned} \mathbf{H} &:= (\mathbf{I} - \mathbf{H}_1)(\mathbf{I} - \mathbf{H}_1)^T - (\mathbf{H}_1 - \mathbf{H}_2)(\mathbf{H}_1 - \mathbf{H}_2)^T + \mathbf{I} \\ &= 2\mathbf{I} - 2\mathbf{H}_1 + \mathbf{H}_2 \end{aligned} \quad (7)$$

it can be shown that the eigenvalues of \mathbf{H} are nonnegative and $f(\mathbf{Z}) = \text{tr}(\mathbf{Z}\mathbf{H}\mathbf{Z}^T)$.

B. Common and Discriminative Dictionaries

The dictionary from solving (1) is optimized for reconstructing the given data. On the other hand, in (5), the dictionary is learned to maximize the classification accuracy through Fisher criterion. Thus, the latter formulation tends to disregard the patterns in the data that do not contribute to classification. In fMRI data analysis, the components that are common across the subject pool (and thus do not contribute to classification) as well as the components that are unique to different subject groups are both important.

To capture both the common and the individual components, \mathbf{D} is split into $\mathbf{D} := [\tilde{\mathbf{D}}, \bar{\mathbf{D}}]$, where $\tilde{\mathbf{D}} \in \mathbb{R}^{V \times \tilde{K}}$ is the common dictionary, which contributes to reconstruction but not as much to classification, and $\bar{\mathbf{D}} \in \mathbb{R}^{V \times \bar{K}}$ is the discriminative dictionary, which is tailored to the classification task. The sparse factor \mathbf{Z} is also split compatibly as $\mathbf{Z} = [\tilde{\mathbf{Z}}^T, \bar{\mathbf{Z}}^T]^T$. Thus, the overall formulation becomes

$$(P1) \quad \min_{\mathbf{D}, \mathbf{Z}} \frac{1}{2} \|\mathbf{X} - \mathbf{D}\mathbf{Z}\|_F^2 + \lambda \|\mathbf{Z}\|_1 + \frac{\mu}{2} f(\tilde{\mathbf{Z}}) + \eta \|\bar{\mathbf{D}}\|_*$$

Input: $\mathbf{X}, \mathbf{D}^{(0)}, \text{MAX_ITER}, \rho > 0$
Output: $\mathbf{D}^{(\text{MAX_ITER})}, \mathbf{Z}^{(\text{MAX_ITER})}$

```

1: For  $\ell = 0, 1, \dots, \text{MAX\_ITER} - 1$ 
2: Compute  $\mathbf{Z}^{(\ell+1)}$  by (8) using, e.g., FISTA
   /* Update  $\tilde{\mathbf{D}}$  with  $\bar{\mathbf{D}}$  fixed */
3: Set  $\tilde{\mathbf{D}} = \tilde{\mathbf{D}}^{(\ell)}$  and  $\tilde{\mathbf{X}} = \mathbf{X} - \tilde{\mathbf{D}}^{(\ell)}(\bar{\mathbf{Z}}^{(\ell+1)})^T$ 
4: Set  $\mathbf{A} := \tilde{\mathbf{Z}}^{(\ell+1)}(\tilde{\mathbf{Z}}^{(\ell+1)})^T$  and  $\mathbf{B} := \tilde{\mathbf{X}}(\tilde{\mathbf{Z}}^{(\ell+1)})^T$ 
5: Repeat
6:   For  $k = 1, 2, \dots, \tilde{K}$ 
7:      $\mathbf{u}_k \leftarrow \frac{1}{A(k,k)}(\mathbf{b}_k - \tilde{\mathbf{D}}\mathbf{a}_k) + \tilde{\mathbf{d}}_k$ 
8:      $\tilde{\mathbf{d}}_k \leftarrow \frac{1}{\max\{\|\mathbf{u}_k\|_2, 1\}} \mathbf{u}_k$ 
9:   Until convergence
   /* Update  $\tilde{\mathbf{D}}$  with  $\bar{\mathbf{D}}$  fixed */
10: Set  $\mathbf{V} = \mathbf{U} = \tilde{\mathbf{D}}^{(\ell)}$  and  $\tilde{\mathbf{X}} = \mathbf{X} - \tilde{\mathbf{D}}(\bar{\mathbf{Z}}^{(\ell+1)})^T$ 
11: Set  $\mathbf{E} = \tilde{\mathbf{X}}(\bar{\mathbf{Z}}^{(\ell+1)})^T$  and  $\mathbf{F} = \bar{\mathbf{Z}}^{(\ell+1)}(\bar{\mathbf{Z}}^{(\ell+1)})^T$ 
12: Repeat
13:   Set  $\bar{\mathbf{E}} = \mathbf{E} + \frac{\rho}{2}(\mathbf{V} - \mathbf{U})$  and  $\bar{\mathbf{F}} = \mathbf{F} + \frac{\rho}{2}\mathbf{I}$ 
14:    $\bar{\mathbf{D}} \leftarrow \arg \min_{\bar{\mathbf{D}} \in \bar{\mathcal{D}}} \text{tr}(\bar{\mathbf{F}}\bar{\mathbf{D}}^T\bar{\mathbf{D}}) - 2\text{tr}(\bar{\mathbf{E}}\bar{\mathbf{D}}^T)$ 
15:    $\mathbf{V} \leftarrow \mathcal{S}_{\eta/\rho}(\bar{\mathbf{D}} + \mathbf{U})$ 
16:    $\mathbf{U} \leftarrow \mathbf{U} + \bar{\mathbf{D}} - \mathbf{V}$ 
17:   Until convergence
18:   Set  $\mathbf{D}^{(\ell+1)} = [\tilde{\mathbf{D}} \ \bar{\mathbf{D}}]$ 
19: End For

```

TABLE I: Algorithm for solving (P1)

where $\|\cdot\|_*$ represents the nuclear norm of a matrix, which is the sum of the singular values. Inspired by the low-rank shared dictionary learning (LRSDDL) in [15], the common dictionary is constrained to have a low rank, which is captured by the last term in (P1). However, unlike [15], $\tilde{\mathbf{D}}$ is *not* further split into per-class dictionaries as $\tilde{\mathbf{D}} := [\tilde{\mathbf{D}}_1, \tilde{\mathbf{D}}_2, \dots, \tilde{\mathbf{D}}_C]$. Note that with the per-class dictionaries, detailed order selection is required and the interpretation of the obtained components can become more involved.

In the context of fMRI data analysis, it is widely recognized that the voxel-wise component maps tend to be sparse, thanks to the fact that different brain regions are responsible for various functions that the brain may be engaged in. Given that the data samples $\{\mathbf{x}_n \in \mathbb{R}^V\}$ represent the brain activations over V voxels, it thus makes sense to perform dictionary learning on \mathbf{X}^T rather than on \mathbf{X} , so that the rows of the sparse coefficient matrix \mathbf{Z} correspond to the component spatial maps. Incorporating this observation, we also consider the following formulation.

$$(P2) \quad \min_{\mathbf{D}, \mathbf{Z}} \frac{1}{2} \|\mathbf{X}^T - \mathbf{D}\mathbf{Z}\|_F^2 + \lambda \|\mathbf{Z}\|_1 + \frac{\mu}{2} f(\tilde{\mathbf{D}}^T)$$

where it is noted that the discriminative features are now the rows of the discriminative dictionary $\tilde{\mathbf{D}}$, which is why $\tilde{\mathbf{D}}^T$ is fed into the Fisher cost.

III. ALGORITHM DERIVATION

A. Algorithm for Problem (P1)

In this section, the algorithms for solving Problems (P1) and (P2) are derived. Starting with (P1), it is first noted that although the formulation in (P1) is not convex over all the variables, it is convex with respect to each block \mathbf{D} or \mathbf{Z} . Thus, one can adopt the block coordinate descent (BCD) method to obtain a locally optimal solution. That is, at iteration $\ell + 1$, with $\mathbf{D}^{(\ell)}$ fixed to its ℓ -th iterate, \mathbf{Z} can be updated by

$$\mathbf{Z}^{(\ell+1)} = \arg \min_{\mathbf{Z}} h(\mathbf{Z}; \mathbf{D}^{(\ell)}) + \lambda \|\mathbf{Z}\|_1 \quad (8)$$

where $h(\mathbf{Z}; \mathbf{D}) := \frac{1}{2} \|\mathbf{X} - \mathbf{D}\mathbf{Z}\|_F^2 + \frac{\mu}{2} f(\tilde{\mathbf{Z}})$. This problem can be solved by various algorithms that can deal with the ℓ_1 -norm-based

Input: $\mathbf{X}, \mathbf{D}^{(0)}, \text{MAX_ITER}$
Output: $\mathbf{D}^{(\text{MAX_ITER})}, \mathbf{Z}^{(\text{MAX_ITER})}$
1: For $\ell = 0, 1, \dots, \text{MAX_ITER} - 1$
2: Solve $\mathbf{Z}^{(\ell+1)} = \arg \min_{\mathbf{Z}} \frac{1}{2} \ \mathbf{X}^T - \mathbf{D}^{(\ell)} \mathbf{Z}\ _F^2 + \lambda \ \mathbf{Z}\ _1$
3: Set $\mathbf{A} := \mathbf{Z}^{(\ell+1)} \mathbf{Z}^{(\ell+1)T}$ and $\mathbf{B} := \mathbf{X}^T \mathbf{Z}^{(\ell+1)T}$
4: Set $\mathbf{D} = \mathbf{D}^{(\ell)}$
5: Repeat
6: For $k = 1, 2, \dots, \bar{K}$
7: $\mathbf{u}_k \leftarrow \frac{1}{A(k,k)} (\mathbf{b}_k - \mathbf{D} \mathbf{a}_k) + \mathbf{d}_k$
8: $\mathbf{d}_k \leftarrow \frac{1}{\max\{\ \mathbf{u}_k\ _2, 1\}} \mathbf{u}_k$
9: For $k = \bar{K} + 1, \dots, K$
10: $\mathbf{u}_k \leftarrow [A(k,k)\mathbf{I} + \mu \mathbf{H}]^{-1} (\mathbf{b}_k - \sum_{j=1, j \neq k}^K \mathbf{d}_j A(k,j))$
11: $\mathbf{d}_k \leftarrow \frac{1}{\max\{\ \mathbf{u}_k\ _2, 1\}} \mathbf{u}_k$
12: Until convergence
13: Set $\mathbf{D}^{(\ell+1)} = \mathbf{D}$
14: End For

TABLE II: Algorithm for solving (P2)

regularizer. In this work, the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) is employed [19]. The FISTA requires as an input the gradient of the differentiable part of the objective $h(\mathbf{Z}; \mathbf{D}^{(\ell)})$, which can be obtained from

$$\frac{\partial h(\mathbf{Z}; \mathbf{D})}{\partial \mathbf{Z}} = \left[\frac{\partial h(\mathbf{Z}; \mathbf{D})}{\partial \mathbf{Z}} \right] = -\mathbf{D}^T (\mathbf{X} - \mathbf{D}\mathbf{Z}) + \mu \begin{bmatrix} \mathbf{0}_{\bar{K} \times N} \\ \tilde{\mathbf{Z}}\mathbf{H} \end{bmatrix} \quad (9)$$

where \mathbf{H} is defined in (7).

Next, updating \mathbf{D} can be done separately for $\tilde{\mathbf{D}}$ and $\bar{\mathbf{D}}$. First, the update for $\tilde{\mathbf{D}}$ is based on

$$\tilde{\mathbf{D}}^{(\ell+1)} = \arg \min_{\tilde{\mathbf{D}} \in \tilde{\mathcal{D}}} \frac{1}{2} \|\tilde{\mathbf{X}} - \tilde{\mathbf{D}} \tilde{\mathbf{Z}}^{(\ell+1)}\|_F^2 \quad (10)$$

$$= \arg \min_{\tilde{\mathbf{D}} \in \tilde{\mathcal{D}}} \text{tr}(\mathbf{A} \tilde{\mathbf{D}}^T \tilde{\mathbf{D}}) - 2\text{tr}(\mathbf{B} \tilde{\mathbf{D}}^T) \quad (11)$$

where $\tilde{\mathbf{X}} := \mathbf{X} - \bar{\mathbf{D}}^{(\ell)} \bar{\mathbf{Z}}^{(\ell+1)}$, $\mathbf{A} := \tilde{\mathbf{Z}}^{(\ell+1)} (\tilde{\mathbf{Z}}^{(\ell+1)})^T$, $\mathbf{B} := \tilde{\mathbf{X}} (\tilde{\mathbf{Z}}^{(\ell+1)})^T$ and $\tilde{\mathcal{D}} := \{\tilde{\mathbf{D}} = [\tilde{\mathbf{d}}_1, \dots, \tilde{\mathbf{d}}_{\bar{K}}] : \|\tilde{\mathbf{d}}_k\|_2 \leq 1, k = 1, \dots, \bar{K}\}$ is the set of column-normalized dictionaries. Problem (11) can be solved again by BCD, where the individual columns in $\tilde{\mathbf{D}}$ constitute the block variables. The algorithm is provided in [7], which is reproduced in lines 3–9 in Table I.

The update for $\bar{\mathbf{D}}$ is then based on

$$\bar{\mathbf{D}}^{(\ell+1)} = \arg \min_{\bar{\mathbf{D}} \in \bar{\mathcal{D}}} \frac{1}{2} \|\bar{\mathbf{X}} - \bar{\mathbf{D}} \bar{\mathbf{Z}}^{(\ell+1)}\|_F^2 + \eta \|\bar{\mathbf{D}}\|_* \quad (12)$$

where $\bar{\mathbf{X}} := \mathbf{X} - \tilde{\mathbf{D}}^{(\ell+1)} \tilde{\mathbf{Z}}^{(\ell+1)}$ and $\bar{\mathcal{D}} := \{\bar{\mathbf{D}} = [\bar{\mathbf{d}}_1, \dots, \bar{\mathbf{d}}_{\bar{K}}] : \|\bar{\mathbf{d}}_k\|_2 \leq 1, k = 1, \dots, \bar{K}\}$. This problem can be solved using the alternating direction method of multipliers (ADMM) [15]. The precise steps are provided in lines 10–17 in Table I. Note that the problem in line 14 is in the same form as (11), so the codes in lines 3–9 can be re-used. In line 15, the shrinkage thresholding operator $\mathcal{S}_\tau(\mathbf{M})$ is defined as

$$\mathcal{S}_\tau(\mathbf{M}) = \mathbf{U}_M \text{diag}(\{(\sigma_i - \tau)^+\}) \mathbf{V}_M^T \quad (13)$$

where $\text{diag}(\{\sigma_i\})$ is the diagonal matrix with the diagonal entries $\{\sigma_i\}$, $(x)^+ := \max\{0, x\}$, and the singular value decomposition of rank- r matrix \mathbf{M} is given by $\mathbf{U}_M \text{diag}(\{\sigma_i\}_{1 \leq i \leq r}) \mathbf{V}_M^T$.

B. Algorithm for Problem (P2)

For formulation (P2), again it is noted that the objective function is convex with respect to either \mathbf{D} or \mathbf{Z} , and thus the BCD method is employed. For updating \mathbf{Z} , one simply needs to solve V LASSO

problems for the individual rows of \mathbf{X} , which can be done readily by FISTA. For updating \mathbf{D} , the relevant sub-problem is

$$\mathbf{D}^{(\ell+1)} := \arg \min_{\mathbf{D} \in \mathcal{D}} \frac{1}{2} \|\mathbf{X}^T - \mathbf{D}\mathbf{Z}^{(\ell+1)}\|_F^2 + \frac{\mu}{2} \text{tr}(\tilde{\mathbf{D}}^T \mathbf{H} \tilde{\mathbf{D}}) \quad (14)$$

which can be solved again by adopting the BCD method with the columns \mathbf{d}_k , $k = 1, 2, \dots, K$ of \mathbf{D} as the block variables. For columns $k = 1, 2, \dots, \bar{K}$, that is, the columns that belong to the common dictionary $\tilde{\mathbf{D}}$, the update rule is the same as that for (11). For columns $k = \bar{K} + 1, \dots, K$, i.e., those that belong to $\bar{\mathbf{D}}$, the update for \mathbf{d}_k now becomes

$$\mathbf{u}_k \leftarrow [A(k,k)\mathbf{I} + \mu \mathbf{H}]^{-1} \left(\mathbf{b}_k - \sum_{j=1, j \neq k}^K \mathbf{d}_j A(k,j) \right) \quad (15)$$

$$\mathbf{d}_k \leftarrow \frac{1}{\max\{\|\mathbf{u}_k\|_2, 1\}} \mathbf{u}_k \quad (16)$$

where $A(k, j)$ is the (k, j) -entry of $\mathbf{A} := \mathbf{Z}^{(\ell+1)} \mathbf{Z}^{(\ell+1)T}$ and \mathbf{b}_k is the k -th column of $\mathbf{B} := \mathbf{X}^T \mathbf{Z}^{(\ell+1)T}$. The overall algorithm is presented in Table II.

IV. EXPERIMENTAL SETUP

A. Data Preprocessing

The datasets used in this study are from the Mind Research Network Clinical Imaging Consortium Collection (publicly available at <http://coins.mrn.org>). These datasets were obtained from 271 subjects, 150 healthy controls and 121 patients with schizophrenia, and a description of the multivariate features used in this study are given in [20]. The Standard stimulus is a 500 Hz tone with a probability of occurrence of 0.82, the Target stimulus is a 1 kHz tone occurring with a 0.09 probability. The Novel stimuli constitute random non-repeating digital noise frequencies played at regular intervals with probabilities of 0.09. The subject is expected to press a button when a Target tone is played. A total of $T = 90$ stimuli are presented to the subject for a period of 200 ms at sporadic intervals. For each subject, the time-course data $\mathbf{Y} \in \mathbb{R}^{V \times T}$ is regressed on to a design matrix $\mathbf{G} \in \mathbb{R}^{T \times P}$, which is the result of a convolution of stimulus onset functions and default SPM hemodynamic response function (HRF) as a part of the General Linear Model (GLM). The regressors are then contrasted between the Target and Standard stimuli resulting in a set of $V = 48,546$ spatial features.

B. DL Parameter Tuning

The available data are randomly partitioned into 120 subjects for training, 60 subjects for validation, and 60 subjects for testing. To avoid adding bias to the learned features, the numbers of controls and the schizophrenic patients are maintained equal. The validation process is repeated 100 times with different partitioning.

The total order K of the dictionary was set to 24, which is the order obtained with ICA [20]. However, the division between the discriminative component order \bar{K} and common component order \tilde{K} must be found, with $\bar{K} + \tilde{K} = K$. These and the remaining parameters were tuned using a grid search. Since the number of parameters is rather large, they were optimized in the order of η, λ, μ and \tilde{K} . The parameters were selected using the average classification accuracy on the validation set.

V. RESULTS

A. Results from ICA

For comparison and benchmark, ICA was first conducted based on [20]. The principal component analysis (PCA) with 24 components was applied to the feature matrix \mathbf{X} , and then the entropy

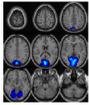
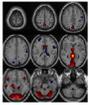
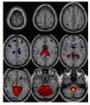
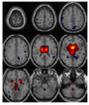
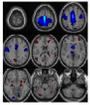
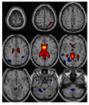
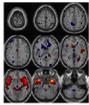
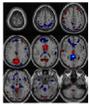
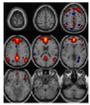
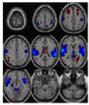
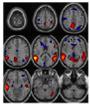
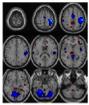
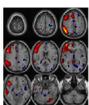
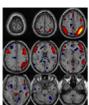
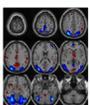
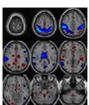
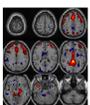
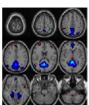
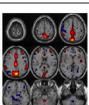
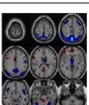
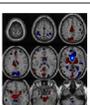
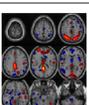
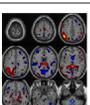
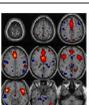
ICA 1	ICA 2	ICA 3 *	ICA 4*	ICA 5*	ICA 6*
					
p=0.3866	p=0.3848	p=0.0015	p=4.34E-06	p=8.73E-05	p=0.0198
ICA 7	ICA 8	ICA 9	ICA 10	ICA 11	ICA 12
					
p=0.5903	p=0.1015	p=0.1866	p=0.325	p=0.1612	p=0.2258
ICA 13	ICA 14	ICA 15*	ICA 16*	ICA 17	ICA 18
					
p=0.0556	p=0.6571	p=6.094E-04	p=0.0447	p=0.1249	p=0.4619
ICA 19	ICA 20	ICA 21	ICA 22	ICA 23	ICA 24
					
p=0.5547	p=0.9862	p=0.531	p=0.2031	p=0.4668	p=0.0803

Fig. 1: Spatial maps obtained from ICA

bound minimization (EBM)-based ICA algorithm [21] was applied to obtain the activation matrix \mathbf{A} and the spatial maps \mathbf{S} . EBM uses a flexible density selection mechanism and has no parameters to select. The signs of the spatial maps were adjusted such that higher activations (red spots in the plots) occur in the controls. To discern the components that are discriminative among the spatial maps found, two-sample t -tests were performed on the activation matrix \mathbf{A} and those that have the p -values less than 0.05 were chosen. In Fig. 1, the spatial maps are depicted, with the statistically significant ones marked in blue.

B. Results from (P1)

Fig. 2 shows the common and discriminative components obtained from solving (P1). The parameter configuration chosen was $\tilde{K} = 13$, $\bar{K} = 11$, $\lambda = 0.005$, $\mu = 0.05$ and $\eta = 0.1$. Based on the two-sample t -test, seven components were found to be statistically significant, which are highlighted in green. The obtained spatial maps were then compared to those from ICA. The components DL 12, DL 15, and DL 18 in Fig. 2 are seen somewhat similar to ICA 19, ICA 5, and ICA 1 in Fig. 1, respectively. The rest four components obtained from (P1) seems to be novel discriminative components that were not seen in the ICA analysis. For example, component DL 14 shows higher activation for healthy controls in cerebrospinal fluid (CSF)/thalamus region, while DL 15 shows activations in the patients group in the motor/sensory motor region.

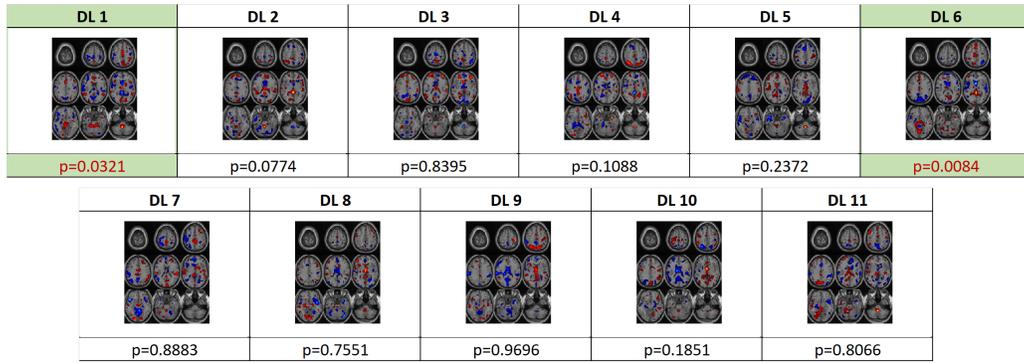
C. Results from (P2)

The grid search for the parameters in (P2) resulted in $\tilde{K} = 18$, $\bar{K} = 6$, $\lambda = 0.001$, and $\mu = 0.05$. The common and discriminative spatial maps from the partition yielding the highest classification accuracy are presented in Fig. 3 along with their p -values. Out of the 18 discriminative components, seven components have the p -values less than 0.05, which are highlighted in orange. There are eight

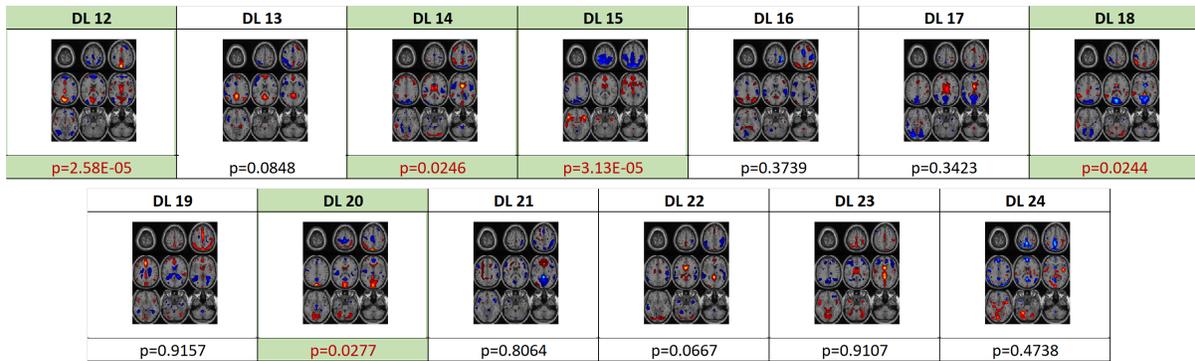
components from Problem (P2) that are analogous the counterparts from ICA. Namely, DL 4, DL 13, DL 14, DL 17, DL 18, DL 19, and DL 24 in Fig. 3 are similar to ICA 13, ICA 3, ICA 4, ICA 5, ICA 16, ICA 15, and ICA 6, in Fig. 1, respectively. It can be seen that (P2) leads not only to more matches between the DL and the ICA components, but also that many of the matching components are statistically significant. That is, (DL 14/ICA 4) indicating the CSF/thalamus region, (DL 18/ICA 16) indicating the parietal region, and (DL 19/ICA 15) indicating the visual and parietal regions are all statistically significant and neurologically relevant for the diagnosis of schizophrenia. Also, there are three components that are uniquely identified as discriminative from DL formulation (P2), which are DL 20, DL 21, and DL 23 in Fig. 3. The component DL 21 has activations in motor and parietal regions which is again deemed to be neurologically relevant. Overall, imposing sparsity in the spatial maps as is done in (P2) seems to lead to more interpretable maps, which also match better with the ICA-based maps.

VI. CONCLUSION

Application of DL techniques to fMRI data analysis was explored to reveal both the common components across the subject population and the individual components that are discriminative of the sub-groups of subjects. Two formulations were proposed, which differed in terms of on which factor the sparsity constraint was imposed, and the corresponding algorithms were derived. When compared with the ICA-based spatial maps, the formulation that imposed sparsity on the spatial activations yielded more matching components. It was also seen that the proposed DL approaches identify a larger number of discriminative components, compared to ICA. In future work, the common and individual components can be obtained by processing multiple modalities jointly, as well as using the resting-state fMRI data.



(a) Common spatial maps



(b) Discriminative spatial maps

Fig. 2: Spatial maps obtained from (P1)



(b) Discriminative spatial maps

Fig. 3: Spatial maps obtained from (P2)

REFERENCES

- [1] S.-J. Kim, V. D. Calhoun, and T. Adali, "Flexible large-scale fMRI analysis: A survey," in *Proc. IEEE Int'l. Conf. Acoustics, Speech and Sig. Process. (ICASSP)*, Mar. 2017, pp. 6319–6323.
- [2] E. S. Finn, X. Shen, D. Scheinost, M. D. Rosenberg, J. Huang, M. M. Chun, X. Papademetris, and R. T. Constable, "Functional connectome fingerprinting: Identifying individuals using patterns of brain connectivity," *Nature Neurosci.*, vol. 18, no. 11, pp. 1664–1671, 2015.
- [3] S. P. Roels, T. Loeys, and B. Moerkerke, "Evaluation of second-level inference in fMRI analysis," *Computat. Intell. Neurosci.*, vol. 2016, pp. 1–22, 2016.
- [4] V. D. Calhoun and T. Adali, "Unmixing fMRI with independent component analysis," *IEEE Eng. Med. Biol. Mag.*, vol. 25, no. 2, pp. 79–90, Mar.-Apr. 2006.
- [5] V. D. Calhoun and T. Adali, "Multisubject independent component analysis of fMRI: A decade of intrinsic networks, default mode, and neurodiagnostic discovery," *IEEE Rev. Biomed. Eng.*, vol. 5, pp. 60–73, 2012.
- [6] T. Adali, M. Anderson, and G.-S. Fu, "Diversity in independent component and vector analyses: Identifiability, algorithms, and applications in medical imaging," *IEEE Signal Proc. Mag.*, vol. 31, no. 3, pp. 18–33, May 2014.
- [7] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res.*, vol. 11, pp. 19–60, Mar. 2010.
- [8] K. Lee, S. Tak, and J. C. Ye, "A data-driven sparse GLM for fMRI analysis using sparse dictionary learning with MDL criterion," *IEEE Trans. Med. Imag.*, vol. 30, no. 5, pp. 1076–1089, May 2011.
- [9] G. Varoquaux, A. Gramfort, F. Pedregosa, V. Michel, and B. Thirion, "Multi-subject dictionary learning to segment an atlas of brain spontaneous activity," in *Proc. Info. Process. Med. Imaging*, Jul. 2011, pp. 562–573.
- [10] A. Mensch, G. Varoquaux, and B. Thirion, "Compressed online dictionary learning for fast resting-state fMRI decomposition," in *Proc. of the IEEE 13th Intl. Symp. Biomed. Imag.*, Prague, Czech Republic, Apr. 2016, pp. 1282–1285.
- [11] A. K. Seghouane and A. Iqbal, "Basis expansions approaches for regularized sequential dictionary learning algorithms with enforced sparsity for fMRI data analysis," *IEEE Trans. Med. Imag.*, vol. 36, no. 9, pp. 1796–1807, Sep. 2017.
- [12] J. Mairal, F. Bach, and J. Ponce, "Task-driven dictionary learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 791–804, Apr. 2012.
- [13] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Fisher discrimination dictionary learning for sparse representation," in *Proc. of IEEE Int'l. Conf. Comput. Vision (ICCV)*, Nov. 2011, pp. 543–550.
- [14] A. Iqbal, K. Seghouane, and T. Adali, "Shared and subject-specific dictionary learning (ShSSDL) algorithm for multisubject fMRI data analysis," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 11, pp. 2519–2528, Nov. 2018.
- [15] T. H. Vu and V. Monga, "Fast low-rank shared dictionary learning for image classification," *IEEE Trans. Imag. Process.*, vol. 26, no. 11, pp. 5160–5175, Nov. 2017.
- [16] D. H. Wolf, B. I. Turetsky, J. Loughhead, M. A. Elliott, R. Pratiwadi, R. E. Gur, and R. C. Gur, "Auditory oddball fMRI in schizophrenia: Association of negative symptoms with regional hypoactivation to novel distractors," *Brain Imag. Behav.*, vol. 2, no. 2, pp. 132–145, 2008.
- [17] I. Tošić and P. Frossard, "Dictionary learning," *IEEE Signal Process. Mag.*, vol. 28, no. 2, pp. 27–38, 2011.
- [18] S. Li and Y. Fu, "Learning robust and discriminative subspace with low-rank constraints," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 11, pp. 2160–2173, Nov. 2016.
- [19] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.
- [20] Y. Levin-Schwartz, V. D. Calhoun, and T. Adali, "Quantifying the interaction and contribution of multiple datasets in fusion: application to the detection of schizophrenia," *IEEE Trans. Med. Imag.*, vol. 36, no. 7, pp. 1385–1395, 2017.
- [21] X.-L. Li and T. Adali, "Independent component analysis by entropy bound minimization," *IEEE Trans. Signal Process.*, vol. 58, no. 10, pp. 5151–5164, Oct. 2010.